DOCUMENT RESUME

ED 453 267 TM 032 804

AUTHOR Snow-Renner, Ravay
TITLE What Is the Promise of Large-Scale Classroom Practice Measures for Informing Us about Equity in Student Opportunities-To-Learn? An Example Using the Colorado TIMSS.
PUB DATE 2001-04-13
NOTE 20p.; Paper presented at the Annual Meeting of the American Educational Research Association (Seattle, WA, April 10-14, 2001).
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Academic Achievement; Access to Education; Accountability; Elementary Education; *Elementary School Students; *Equal Education; *Measurement Techniques
IDENTIFIERS Colorado; Large Scale Programs; *Opportunity to Learn; *Third International Mathematics and Science Study

ABSTRACT

Academic achievement and opportunity to learn (OTL) were studied using data from the 1995 Third International Mathematics and Science Study (TIMSS) for Colorado students at the elementary level. The study used a comprehensive definition of OTL that included content coverage, curricular focus, duration of instruction, and instructional strategies. The implications for using large-scale measures to indicate how fairly educational opportunities were distributed were studied in a context of comparative accountability measures. Data were from 104 third and fourth grade classes in 50 schools, for a total of 2,163 students. The data patterns indicate that a survey measure of OTL can provide some reasonable information about the nature and distribution of instructional practices in different classrooms. This information is at a relatively "broad brush" level of specificity, but it does support the idea that some classes have access to content topics more than others do. Large scale measures of OTL have promise for providing general information about students' experiences in the classroom, but they are not technically sound enough to provide rigorous data in an era of high-stakes accountability. (Contains 5 tables and 46 references.) (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 453 267

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
R. Snow-Renner
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
1

What is the promise of large-scale classroom practice measures for informing us about equity in student opportunities-to-learn? An example using the Colorado TIMSS

Ravay Snow-Renner

Mid-Continent Research for Education and Learning (McREL)
2550 South Parker Road, Suite 500
Aurora, CO 80014

Not for citation without the express permission of the author

Annual Meeting of the American Educational Research Association
Seattle, WA
Session 39.38

TM032804

April 13, 2001

BEST COPY AVAILABLE

What is the promise of large-scale classroom practice measures for informing us about equity in student opportunities-to-learn? An example using the Colorado TIMSS

Educational equity, sometimes phrased as "equality of educational opportunity," has been a consistent theme in American education reform since *Brown vs. Board of Education* in 1954. Liberals and conservatives tend to agree on a general level about the importance of social justice in the schoools, and have incorporated language addressing equity concerns into the rhetoric of recent education reforms about excellence for all students. However, when one considers equity as "a principle that dictates the distribution of educational resources" (Howe, 1989, p. 326), its meanings have been interpreted very differently according to underlying political and philosophical beliefs. These differing interpretations of equity have led to vast differences in understanding its implications for school practice and policy, as well as how to structure research to demonstrate equity (e.g., Guiton & Oakes, 1995).

In the standards reforms, issues of equity have most directly been addressed through the concept of opportunity-to-learn (OTL). As a broad concept, OTL makes intuitive sense; it addresses whether students have had the opportunity to study a particular topic or to learn how to solve a particular type of problem presented by an assessment (McDonnell, 1995). OTL emerged during the national policy debate on standards and testing as "school delivery standards," or OTL standards, which were to be developed collectively by states for each state to select from in determining criteria for assessing the capacities and performance of its schools (National Council on Education Standards and Testing, 1992, p. 13). It was thought that incorporating OTL into the reform would help provide more equitable learning opportunities for students before they were held accountable for achievement (Conference Report, 1994; McLaughlin & Shepard, 1995). However, the redistributive potential of OTL standards resulted in hot bipartisan debate in the mid-1990s about their use and definition, with the result that OTL was stripped from legislation and has all but disappeared from the policy arena.

Without any provision made for student OTL, current realizations of standards-based reforms do not take current system inequities into account. States are rushing to implement high-stakes accountability measures based largely on student achievement scores. In some cases, data about the school and district context are also considered in connection with achievement data (Olson, 1998; Education Week, 2001), but in general, these measures are relatively far-removed from the classroom context. In the interest of fairness to students, teachers, and schools, before achievement data are used to inform policy decisions, concerted efforts should be made to examine the context of the learning environment (American Educational Research Association [AERA], 1999).

Within the current policy context which bases accountability measures mainly on achievement, lawsuits against schools, districts, or states that fail to provide adequate OTL (O'Day & Smith, 1992; Welner & Oakes, 1996) are no longer a distant threat but, in fact, are already occurring (e.g.,*GI Forum v. Texas Education Agency*, a 1999 extension of the 1981 *Debra P. vs. Turlington* case). The approach recently taken by Texas in successfully defending its graduation measure was to establish that due process had been consistently met by the state testing office. However, another possibility is to examine the measures and ways by which we can best determine whether OTL is being provided for students in individual classrooms, particularly given classroom-level variations in OTL introduced by institutional structures like ability tracking and school-level variations related to funding disparities.

This paper describes findings from a study of achievement and OTL based on the 1995 Colorado TIMSS data at the elementary level. The study used a comprehensive definition of OTL that included content coverage, curricular focus, duration of instruction, and instructional strategies. This paper addresses the findings about classroom-level OTL in light of the implications for using large-scale measures to inform us about how fairly educational opportunities are distributed within a context of comparative accountability measures.

Background

The research and policy context of OTL

McDonnell (1995) reports that OTL originated as a technical research term designed to increase the validity of cross-national comparisons of student mathematics achievement conducted by the IEA. IEA researchers realized that they needed to take national curricular differences into account when conducting comparative studies of international achievement. Therefore, OTL, which addresses whether students have had the opportunities to learn the content on which they will be tested, was introduced as part of the First International Mathematics Survey in the early 1960's, but was refined by the second administration of the study (the Second International Mathematics Study, or SIMS), conducted between 1976 and 1982. During SIMS, teachers were surveyed about whether they had taught the content needed to respond to specific items administered on the test, as well as about their general goals, beliefs, instructional strategies, and professional preparation (McDonnell 1995; Schmidt & McKnight 1995).

By the third administration of the study (the Third International Mathematics and Science Study, or TIMSS), in 1995, the OTL measure was further refined, using teacher surveys that addressed a broader range of OTL than previously. Item-specific coverage questions were limited to a subsample of test items at the seventh and eighth grade levels, while additional general questions about content coverage and instructional strategies were incorporated. Alternate data sources were incorporated into the TIMSS design to provide validity information about the OTL surveys. These included a videotape study and curriculum and textbook analyses (Schmidt & McKnight, 1995).

In the mid-1980's, after the administration of SIMS, IEA findings about OTL began to influence the development of curriculum and process indicators, particularly in the areas of mathematics and science. Recommendations were made to include such information in the educational indicator data collected regularly by states and the federal government (McDonnell, 1995). Apart from IEA efforts, the research about classroom processes tended to be small-scale and based on unrepresentative samples or else were focused on traditional resource inputs that did not capture variation at the classroom level (Brewer & Stacz 1996). A variety of reports (Blank, 1993; Guiton & Burstein 1993; Raizen & Jones 1985; Stecher, 1992) indicated the need to develop and collect data about school and classroom processes on a broad scale.

OTL emerged during the national policy debate on standards and testing in the early 1990's as "school delivery standards." As envisioned by the National Council on Education Standards and Testing (NCEST), a fully-realized standards-based educational system included:

- *content standards* that describe the knowledge, skills, and other understandings that schools should teach in order for students to attain high levels of competency in challenging subject matter;
- *student performance standards* that define various levels of competence in the challenging subject matter set out in the content standards
- *school delivery standards* developed by the states collectively from which each state could select the criteria that it finds useful for the purpose of assessing a schools' capacity and performance; and
- *system performance standards* that provide evidence about the success of schools, local school systems, states, and the Nation in bringing all students, leaving no one behind , to high performance standards
  (NCEST, 1992, p. 13)

Later, the policy terminology changed from "delivery standards" to "opportunity-to-learn standards."

OTL standards proved to be contentious during early conversations about standards reform. Advocates envisioned them as a way to hold policy makers accountable for providing adequate opportunities for learning to students traditionally underserved by the educational system (O'Day & Smith, 1992). Without OTL standards, it was feared that students would unfairly assume all the consequences for the current inequities in curriculum, instruction, and learning environments, rather than schools and systems. However, others raised concerns about whether OTL standards were appropriate vehicles for addressing the equity and quality problems of education (McLaughlin & Shepard, 1995; Traiman, 1993). In general, the policy debate over OTL standards has tended to center on how such standards would be defined, what their purpose and use should be, when they should be developed during the implementation process, and what the role of the federal government should be (Traiman, 1993, p. 12).

The intensity of conversations about OTL has abated in the policy realm since the mid-90's. OTL standards were incorporated into law as part of the Goals 2000 legislation in 1994. They were defined as:

> the criteria for, and the basis of, assessing the sufficiency or quality of the resources, practices, and conditions necessary at each level of the education system (schools, local educational agencies, and States) to provide all students with an opportunity to learn the material in voluntary national content standards or State content standards.
> (Goals 2000: Educate America Act, 1994, Section 3)

However, McDonnell (1995) notes that, as policy, OTL standards thus operationalized fall somewhere between a hortatory policy (e.g., indicators of OTL provide information that is available for people to use voluntarily in guiding practice) and a weak inducement (e.g., by developing OTL standards, states may receive some small amounts of federal funding). OTL standards were repealed from the Goals 2000 legislation in 1996, and since that time, have remained largely outside the realm of policy discussions around standards, although the topic was renewed somewhat in the context of the Clinton administration's ill-fated 1997 proposal for national student achievement testing in reading and mathematics.

The result of these earlier policy debates over OTL is that, while the education system is now geared to standards and assessments, it lacks even a formal provision for OTL standards. Further, assessments are increasingly taking the form of high-stakes accountability measures. Education Week, in a recent analysis of accountability reform in the United States, notes that 49 of the 50 states have state standards and 50 have state assessments to measure student achievement. Forty-five states generate school report cards, and 27 of those rate school performance primarily on student test scores (Education Week, 2001, p. 15).

Introduction to TIMSS

If OTL measures are described in terms of cars, the TIMSS (1995) and its more recent re-administration measure, the TIMSS-R, may well be considered "Cadillacs." Consisting both of comprehensive, broadly-studied survey measures of teacher, school, and student background information, and large-scale comparative measures of achievement in mathematics and science, the TIMSS represents the culmination of almost 40 years of research that has gone toward refining and validating subsequent measures of OTL. Additionally, the survey materials have been widely studied, with triangulation studies conducted, including curriculum analyses and a videotape analysis (Schmidt & McKnight, 1995).

The 1995 TIMSS was a large-scale comparative study of national education systems in 45 different countries. TIMSS researchers examined mathematics and science curricula, instructional practices, and school and social factors, as well as conducting achievement testing of students at three different instructional levels. Data were collected from representative documents laying out official curricular intentions and plans, mathematics and science textbooks, and researchers also searched K-12

textbook series for selected "in-depth" topics (subareas within the broader subject matter). In six countries, TIMSS conducted classroom observations, teacher interviewing, and videotaping. (Martin, 1996; Schmidt, McKnight, & Raizen, 1997). In addition to the international administration of TIMSS in 1995, three U.S. states opted to administer the TIMSS to selected student, teacher, and administrator populations at three different instructional levels. Colorado was one of these states, and participated in the TIMSS for age 9 students (sampled in grades 3 and 4).

TIMSS has been organized to measure curriculum as a "broad explanatory factor underlying student achievement" and that curriculum is considered to have three manifestations, originally conceived for the IEA's Second International Mathematics Study (SIMS):

1. What society would like to see taught (or the *intended curriculum*);
2. What is actually taught in the classroom (or the *implemented curriculum)*; and
3. What students learn (the *attained curriculum*). (Martin, 1996)

While this three-pronged approach does focus attention on three different interpretations of curriculum, the underlying concept unifying these approaches, according to key TIMSS researchers, is based on the *provision of educational opportunities to students* (Martin, 1996; Schmidt & McKnight, 1995) at different levels of the system.

One role of TIMSS data was to provide comparative achievement information relative to Goals 2000;TIMSS was the comparative measure designed to track progress toward Goal 5, the assertion that the U. S. would be "first in the world" in mathematics and science. The TIMSS findings were disappointing for policy makers. The data indicated that U.S. fourth graders were closest to that goal in 1995; they ranked 8th of 25 participating countries, while eighth graders ranked 20th of 41 participating countries and twelfth graders performed among the lowest of the 21 TIMSS countries participating at that education level. U.S. students at both eighth and twelfth grade levels performed below the international averages (U.S. Department of Education, National Center for Education Statistics, 1996, 1997, 1998). These are the TIMSS findings most widely bruited, rather than information about U.S. curriculum.

In addition to its comprehensive nature and public presence in the research community, policy makers consider the TIMSS data to be very important. Its findings have received an unprecedented amount of public attention, largely due to their emphasis on international comparisons of mathematics and science achievement. TIMSS' broadest findings, which involve ranking of nations by grade levels in international achievement, have been widely cited by policy makers, researchers, and other groups involved in public education as an example of the U.S.' competitive "vulnerability"relative to other nations (Campbell & Clewell, 1999). Frequently, the TIMSS achievement results are used as justification for a variety of proposed policy changes in education to reduce this perceived international vulnerability (Wolf, 1998). However, the potential of the TIMSS measures to inform the field about the state of the art in OTL indicators–their feasibility and limitations--is also important, particularly given the current context of education policy.

Indicators of OTL in the research

This study focuses on three overlapping purposes from OTL indicators as described in the literature; to use for monitoring classroom practices and student learning opportunities in the interest of equity–particularly in terms of determining the validity of comparing performance across different schools or districts, to provide contextual information for informing accountability policy, and to examine in relationship to student achievement–in terms of what works.

The primary purpose described here is for measuring and monitoring student learning activities in the interest of equity (Guiton & Oakes, 1995; Herman, Klein, & Wakai, 1996; Howe, 1989; Welner & Oakes, 1996). O'Day and Smith (1992) advocate a variety of ways in which OTL standards may inform the distribution of student opportunities. Additionally, OTL information was originated and has more

recently been described as a way of validating comparisons of student performance (Schmidt & McKnight, 1995; Wiley & Yoon, 1995;) in the interest of fairness. Finally, the use of OTL indicators to measure equity has also been recommended specifically within the context of informing accountability procedures, with the option of high stakes use of the data for schools (Herman & Klein, 1997).

Related to the first purpose, the second purpose of OTL indicators is more specifically focused on providing contextual information to inform accountability processes (O'Day & Smith, 1992; Schmidt & McKnight, 1995; Muthen, Huang, Jo, Khoo, Goff, Novak, & Shih, 1995). This could take the form of hard accountability policy, in which schools are held accountable for the opportunities provided, or softer policies, such as providing contextualizing information around student achievement–a distinction between different *stakes* of test results, or "the importance of the results of testing programs for individuals, institutions, or groups" (AERA, 1999, p. 139). In the literature, the issue of purposes for OTL indicators tends to become confused with the issue of stakes. Porter (1995), in examining OTL in the context of standards-based reform, distinguishes between monitoring OTL indicators for accountability (e.g., comparative and high stakes) purposes and for monitoring for school improvement purposes (rather more low-stakes), while McDonnell provides a four-point continuum of stakes for tests, depending on whether the results are used to develop policy *as* mandates, hortatory policy, inducements, or as part of a strategy to build local capacity (McDonnell, 1995, p. 313-314).

The third purpose for examining OTL indicators is for providing information about instructional factors that facilitate or hinder student achievement, or as "what works." One criterion for the development of OTL standards, as debated early in the standards movement, was that they should be relevant to achievement, and a considerable body of the research focuses on this role (Brewer & Stacz, 1996; Hanushek, 1997; Porter, 1995). In addition to looking at what works as a general purpose, a variety of researchers have focused primarily on technical issues of OTL measurement, largely as it relates to student achievement. In general, these findings have been mixed, largely because of a variety of approaches to OTL measurement and a variety of different achievement measures.

Howe (1989) provides an argument for using the distribution of student achievement outcomes as an indicator of equity. He begins by critiquing the argument that unequal outcomes in education are a result of choices on the parts of students whether or not to take advantage of opportunities, and argues that outcomes rather than choices should be used as ways to measure student equality of opportunity. He notes, however, that these outcomes should be qualified by "morally irrelevant characteristics" relating to achievement. Such characteristics might include race, while morally relevant characteristics might be characterized as academic ability. Indicators of OTL may provide an empirical tool for defining morally relevant characteristics of instruction by examining the specific instructional processes that relate to student achievement. At any rate, in the interest of fairness and abundant studies that student opportunities vary by school (Kozol, 1991); track placement (Oakes, 1985), and individual classes (Gamoran, 1987), the study of classroom processes is important to help determine the equitability of student opportunities.

The study

This paper presents findings about classroom-level indicators of OTL as reported by elementary mathematics teachers in response to TIMSS survey items. The study was organized in three different stages: exploration of student achievement data in mathematics, exploration of OTL patterns, particularly in regard to their distribution, and the examination of how these variables were related to each other.

Sampling

TIMSS achievement and OTL data were gathered in spring, 1995, from a representative sample of Colorado nine-year-old students enrolled in public schools in grades 3 and 4. Based on the student sample, information about instructional and school-level processes was gathered from their teachers and

administrators as well (Martin, 1996; WESTAT, 1997). A two-stage sampling design of classes within schools was used; 50 schools were randomly sampled, with a random sample of third and fourth grade classrooms sampled within schools. All students in the sampled classrooms took the TIMSS survey, while teachers were selected to complete surveys addressing OTL based on whether they taught assessed students mathematics and/or science.

For the purposes of this study, I used a selective subsample of the larger WESTAT data, combining data from teacher and student databases and utilizing information only from those classes with both student achievement and teacher OTL data. Data used are from 104 third and fourth grade classes in 50 schools--47 third-grade classrooms, and 57 fourth-grade classrooms--and account for a total of 2,163 third and fourth grade students who took the TIMSS achievement test.

Stage 1: Operationalization of achievement and emergent data patterns

Because of the selective subsampling procedures I used to develop this data set, weighted plausible values, which are used to report TIMSS achievement results at state and national levels, were not used in this study. Further, as a significant aspect of classroom-level OTL has been operationalized in terms of content coverage (Brewer & Stacz, 1996), it was important to separate global mathematics achievement scores into specific subtopics. Groups of achievement items on the TIMSS measure mapped specifically onto six different subtopics of mathematics achievement: Whole numbers; Fractions and proportionality; Measurement, estimation, and number sense; Data representation, analysis, and probability; Geometry; and Patterns, relations, and functions.

To generate student and classroom-level data, I used the raw data to recalculate individual student response scores, based on the items students were actually assessed on within the matrix sampling scheme, which utilized eight different text booklets. This information was cross-referenced to each of the six mathematics subtopics listed above to develop subscale variables that represented the percentage correct of items on each particular subtopic. After checking classroom-level distributions, I then aggregated these scores to the classroom level, and analyzed them by grade to clarify the distribution of achievement patterns. A summary of descriptive statistics for 3rd and 4th grade classroom achievement is provided in Table 1.

Table 1. Student mathematics achievement by subscale

Subscale Content	*Grade 3 (N = 45)*			*Grade 4 (N = 51)*		
	Range	*M*	*S*	*Range*	*M*	*S*
Whole Numbers	23% - 81%	48.7%	.1337	32 - 82%	61.5%	.1063
Fractions and Proportionality	15% - 59%	35.6%	.0951	18% - 66%	44.9%	.1062
Measurement, Estimation, and Number Sense	19% - 58%	38%	.0914	17% - 72%	46.9%	.1174
Data Representation, Analysis, and Probability	18% - 87%	53.8%	.1452	31% - 86%	65.8%	.1291
Geometry	19% - 66%	45.9%	.1053	22% - 79%	56.8%	.1320
Patterns, Relations, and Functions	13% - 72%	42.2%	.1390	23% - 82%	55.2%	.1329

Because these findings were based on raw data, rather than IRT-generated plausible values scores, they should be interpreted with caution.

To explore their utility and limitations, therefore, I examined their reliability and validity. Subscale reliability varied by content area and booklet, but, with two exceptions (*Geometry* and *Patterns*), demonstrated reasonable levels of consistency reliability, with the lowest reliability estimated for the *Measurement , Estimation, and Number Sense* subscale. The *Geometry* and *Patterns* subscales were omitted from the remainder of the study. To explore subscale validity, I compared the classroom-level scores with overall achievement patterns in the weighted statewide data, as reported by Voelkl (1998) and illustrated in Table 2.

Table 2. Comparisons of weighted and unweighted subscale achievement

Data Source	Grade	Whole numbers (25 items)	Fractions & Proportionality (21 items)	Measurement estimation, & number sense (20 items)	Data representation, analysis, & probability (12 items)
Voelkl, 1998 (weighted N = 103,776 students)	3	54	33	37	53
	4	67	45	48	68
Study sample, 1999 (N = 96 classrooms)	*3*	*48.7%*	*35.6%*	*38%*	*53.8%*
	4	*61.5%*	*44.9%*	*46.9%*	*65.8%*

(Weighted data derived from Voelkl, 1998, pp. 38, 39)

Comparisons of the unweighted data with Colorado weighted achievement scores indicated that subscales captured broadly similar patterns, in terms of relative strengths and weaknesses. Students at both grade levels performed most strongly in the areas measuring data analysis skills and most poorly on fractions and measurement-related items. In comparing the weighted data with subscale-generated data, one descriptive pattern that emerges is the relative consistency in patterns of strong and weak performance, although the actual percent correct is different for each data set.

A final exploration of the technical quality of subscales involved examining differences in achievement by grade level. As other researchers exploring the TIMSS and using plausible values have found systematic differences in performance between upper and lower-grade students (Voelkl, 1998; Schmidt, McKnight, Cogan, Jakwerth, & Houang, 1999), similar patterns in these subscales provide further evidence of score validity, in terms of the subscale measures' sensitivity to instructional effects. A Multivariate Analysis of Variance (MANOVA) of all six achievement subscales indicated that fourth grade classrooms performed significantly better than third grade classrooms on *all* mathematics subscale measures ($p < .001$ for all subscales). This shows that, in addition to being broadly responsive to curriculum-specific elements of achievement, subscale scores are also relatively sensitive to grade level differences that might be explained by a general "instructional" effect, broadly defined.

Stage 2: Operationalization of OTL and emergent data patterns

In general, classroom-level indicators of OTL fall into three different categories; *content coverage*, *instructional strategies*, and *classroom resources* (Brewer & Stacz, 1996). For this study, OTL data from a WESTAT-provided database of individual teacher responses to an extensive teacher background survey were used that addressed *content* and *instructional strategies*. Content was measured in terms of:

- curricular focus, operationalized by the number of topics covered by teachers during the school year previous to the administration of the TIMSS; and
- topic coverage or non-coverage

These aspects of OTL were measured in the TIMSS using a block of variables addressing frequency of coverage of 36 different specific mathematics topics and subtopics. An additive focus variable was created by adding up the total number of topics reported covered by a given teacher. The higher this number, the *less* curricular focus described in the classroom. This is a rough parallel, although not a replication, of the way that curricular focus was studied by the U.S. TIMSS Center (Schmidt, McKnight,& Raizen, 1997).

The coverage variables were developed so that they mapped onto the achievement subscales, based on the judgment of mathematics content experts. Because achievement and OTL domains were framed at different levels of specificity, this was quite a feat. If the experts did not consider any particular OTL topic to map appropriately to any student performance area at the elementary level, these variables were omitted from further analysis. Further, if OTL topics from the teacher questionnaire mapped onto more than one achievement area, they were counted as a component of all relevant areas.

In addition to access to ambitious content, another aspect of OTL is access to *ambitious instruction*, which is sometimes typified by strategies advocated by the National Council of Teachers of Mathematics (NCTM). Instructional strategies in this study were operationalized using six variables. On these variables, teachers reported the frequency with which they required their students to participate in key learning activities, including:

- explaining the reasoning behind an idea;
- representing/analyzing relationships using tables, charts, or graphs;
- working on problems with no immediately obvious method of solution;
- practicing computational skills;
- using computers to solve exercises or problems; and
- writing equations to represent relationships.

Some, but not all, of these six different activities can be interpreted in terms of their relative orientation to *reform-oriented*, constructivist practices in mathematics, as exemplified by the NCTM's Curriculum and Evaluation Standards for School Mathematics (1989), or more *traditional* practices that reflect the influence of drill-and-practice or other more teacher-centered types of pedagogy. Initial patterns were explored for all six variables; the last two were excluded from the final analysis linking OTL with achievement based on findings from an exploratory factor analysis. This factor analysis demonstrated empirically that the first three items load onto a reform factor, consistent with the tenets of the NCTM documents (e.g., emphasizing reasoning, multiple methods of analysis, and problem-solving). The computation item loaded on a more traditional factor.

Content-oriented OTL findings

In terms of curricular focus, the distribution of topics reported taught by teachers was lower in general for third grade teachers than for fourth grade teachers. The minimum number of topics for third grade classrooms was 8 (compared to 9 for fourth-grade classrooms) and the maximums were 29 and 35 topics for third and fourth grades, respectively. On average, third grade teachers reported covering 16.39 mathematics topics, compared with 22.02 topics for fourth grade teachers. These data indicate considerable variation across classrooms. Some teachers report teaching almost all topics listed, while others report covering much fewer, with a systematic pattern of variation by grade level-- fourth grade teachers teaching significantly more topics (e.g., reporting *less* focus) than third grade teachers, as verified by an ANOVA of mean topics covered ($df = 1, 87$, $F = 21.285$, $p<.001$).

These findings do not validate international reports of curricular focus on the TIMSS. One of the hypotheses put forth by international researchers is that poor international performance by U.S. students is partially due to an unfocused curriculum (Schmidt, et al., 1997). But these data patterns indicate that, at least on the macroscopic level, the positive relationship hypothesized between *focus* and *achievement* by

international researchers is not apparent here. We can already see that fourth grade teachers report covering more topics than third grade teachers, (thus demonstrating *less* curricular focus) and we know from our prior analyses of subscale achievement that fourth grade classes have significantly higher achievement levels.

Topic coverage was examined using frequencies of teachers who reported coverage and noncoverage by grade level across the list of topics, and differences by grade level were explored using chi-square tests of independence for all topics. A general summary of findings for the four subscale content categories of *Whole numbers, Fractions and Proportionality, Measurement, Estimation, and Number Sense*, and *Data Representation, Analysis, and Probability* are provided here.

Whole numbers , *Measurement*, and *Data representation* topics are relatively broadly covered to teachers across grade levels.

- *Whole numbers* topics are taught by large proportions of teachers at each grade level–across all four subtopics, more than 80% of all teachers across both grades reported coverage. No significant differences in coverage took place between grade levels.

- In the area of *Measurement*, which shares several subtopics with *Whole numbers*, teachers also report broad coverage across grade levels, particularly on skills addressing estimation and number sense and measurement units and processes, with few differences in coverage between third and fourth-grade teachers. However, in the topic of number theory, significantly more (76.4%) fourth grade teachers report coverage than do third grade teachers (approximately 40%) ($p<.001$)

- In the area of *Data representation*, teachers across grade levels report relatively broad coverage on data representation and statistics, although only about 50% of teachers at each grade level report that they cover the topic of probability. There are no significant differences in coverage by grade level.

In the area of *Fractions*, however, considerably more variation in coverage is apparent. First of all, there is considerably more evidence of grade-level articulation in fractions than for whole numbers. All fractions topics save one that addresses the meaning, representation and uses of common fractions, are covered by significantly more fourth grade teachers than third grade teachers ($p<.05$). Additionally, considerably fewer teachers at both grade levels reported coverage on fractions topics than on whole numbers–particularly for topics addressing more complex content. These findings are congruent with other studies indicating that, in the U.S., decimal fractions topics are introduced into the curriculum at fourth grade (Schmidt, et al., 1999).

These data patterns are consistent with other studies showing that topic coverage in mathematics tends to emphasize relatively low-level skills. In this data set, teachers most broadly emphasize topic coverage that focuses on *Whole numbers* and the simpler aspects of *Fractions*. Most teachers at each grade level address topics related to whole numbers. However, there is slightly less emphasis on whole numbers at grade four than at grade three; fewer teachers at grade four report general coverage of whole numbers. This may be explained partially by the relatively greater emphasis on fractions-related topics at the fourth grade.

Additionally, the data patterns highlight some curriculum articulation issues; fourth grade teachers are significantly more likely to address aspects of *Fractions* than third grade teachers, most of whom do not report coverage of fractions content beyond meanings of common fractions. This is particularly true for more complex content. However, fourth grade coverage of fractions is broadly characterized by an emphasis on lower-level content, particularly on properties of common fractions. Aspects of decimal fractions and other, more complex content, are left largely unaddressed in most fourth grade classrooms.

Student learning activities–OTL findings

In examining how student learning activities were organized in different classes, I first conducted a MANOVA to examine whether teachers differed significantly by grade level in their use of these strategies. The MANOVA indicated that there were no statistically significant differences; therefore, I explored descriptive statistics and frequencies on the six variables. From these, I could see the following things:

- The predominant learning activity in these classrooms is the *practice of computational skills*, a practice typically associated in the reform literature with *traditional* instruction.
- The second most predominant practice is *students explaining the reasoning behind their ideas*, a technique commonly associated with reform recommendations.
- Finally, student *use of computers to solve exercises or problems* is the least-frequently reported learning activity.

Examination of frequencies provided more information. In addition to being the predominant practice, *practicing computation* is a learning activity that takes place *every class lesson* in more than 17% of these teachers' classes. No teachers report that this *never or almost never* happens in their classrooms. Similarly, *students explaining their reasoning* is addressed by almost all classrooms; only 1 teacher (approximately 1%) says this *never or almost never* occurs, and most (almost two-thirds) say that this activity is required of students during *most lessons* or *every lesson*.

Similarly, the predominant pattern of task assignment for most students is that of access to tasks with fairly closed-ended solutions. Only about 9% of teachers say that their students *work on problems without immediate solutions* as frequently as *most lessons* or *every lesson*. Additionally, the *use of charts and graphs to represent relationships* is not very frequent; almost 17% of teachers say this is *never or almost never* done in their classrooms, although 70% say this happens during *some lessons*. Less than 10% of the teachers' students do this *most lessons* or *every lesson*.

Finally, the *use of computers to solve exercises or problems* is uneven across classrooms; almost one-half of the teachers say that their students *never or almost never* do this. Another, similar proportion (46.1%) reports that their students do this *some lessons*. Only approximately 3% of the teachers require students to work on computers *most lessons* and *no* teachers report that this happens *every lesson*. It is likely that differential access to computers and technology in the classroom, as well as varying levels of teacher training in instructional technology use, are related to this data pattern–an important equity issue.

These findings indicate that learning activities in these classrooms vary considerably and that teachers report an eclectic variety of instructional practices, similar to findings from the Minnesota TIMSS (Lawrenz & Huffman, 1998). The most common learning activity is the traditional practice of computational skills, the data indicate that most student learning opportunities are based on relatively traditional, closed-ended problems, and almost 17% of these classrooms do not usually use charts and graphs to represent relationships--relatively traditional forms of instruction. However, on the other hand, students frequently explain the reasoning behind a mathematical idea; almost 2/3 of the teachers say this happens *most* or *every lesson*–and 70% of the teachers report that charts or graphs are used during *some lessons*–more reform-oriented activities. This eclecticism is echoed in other, more ethnographic studies about teacher use of reform practices (Cohen, 1990; Spillane & Zeuli, 1999).

Stage 3: Examining relations between achievement and OTL

While the emphasis of this study is primarily on the feasibility of using large-scale measures of OTL to monitor equity issues, an important consideration, particularly in the context of current achievement-based accountability reforms, is how well the OTL indicators studied link to student achievement. This is a key area of interest to policy makers and educators alike. To examine this

relationship at the classroom level, I conducted an initial bivariate correlation between OTL and achievement measures. OTL variables included: curricular focus, topic coverage (in the areas of whole numbers, fractions, measurement, and data analysis), and four student learning activities (explaining the reasoning behind an idea; representing/analyzing relationships using tables, charts, or graphs; working on problems with no immediately obvious method of solution; and practicing computational skills) as selected through a factor analysis. They were correlated with the four classroom-level achievement subscale scores that had high levels of reliability (whole numbers, fractions, measurement, and data analysis). This analysis was based on a logical hypothesis; it is reasonable to expect that students who have received instruction on a topic (particularly a complex topic like decimal fractions) should, on average, score higher on a measure of topic-specific knowledge than students who had not received such instruction. Additionally, this approach made it possible for me to study the hypothesis of international researchers about the relationship between curricular focus and achievement by examining data patterns internal to the Colorado data set. Table 3 illustrates the significant correlations found in this study between OTL and achievement variables. The OTL variables for which no significant correlations were found for either grade on any of the four achievement measures are not included in the chart (e.g., coverage of data analysis topics, students' explanation of their reasoning, and student use of charts, tables, or graphs to represent or analyze mathematical relationships).

Table 3. Significant correlations between OTL and achievement variables

OTL Variable	*Grade*	*N*	*Whole Numbers Score*	*Fractions Score*	*Measure-ment Score*	*Data Analysis Score*
Curricular Focus (number of topics covered in mathematics class)						
Number of mathematics topics taught	3	36	**.428****	**.422***	**.459****	**.372***
	4	49	**.359***	**.321***	**.329***	**.289***
Topic Coverage on:						
Whole Numbers Topics	3	45	.268	.251	**.374***	.234
	4	51	-.116	-.121	-.111	.014
Fractions topics	3	45	**.336***	**.439****	**.365***	.252
	4	51	-.062	-.007	-.079	-.015
Measurement topics	3	45	.286	.282	**.388***	**.308***
	4	51	-.110	-.122	-.097	-.011
Student learning activities						
Practice of computational skills	3	45	.148	.223	.114	.188
	4	50	**-.353***	-.219	-.305	**-.353***
Doing problems with no obvious method of solution	3	45	.232	.122	.174	.183
	4	50	**.293***	**.312***	.269	**.400****

* p<.05
** p<.001

The only variable that correlates significantly and positively across grade levels with all four achievement subscales is the *curricular focus* variable (the number of mathematics topics taught). However, given the way I structured the variable, one would expect an inverse relation. Therefore, the relationship that Schmidt, et al. (1999) hypothesize between focus and student achievement is not

apparent in these data. In fact, a relationship of the *opposite* direction is seen for both grade levels, in which students tend to perform significantly better on all subscales when teachers report coverage of larger numbers of mathematics topics.

For the other variables, correlations are inconsistent by grade level. For some subscales, I found moderate correlations with OTL items that related to topic coverage. However, these patterns were the case *only for third grade classes.* No fourth grade classes showed any significant relationships between achievement and topic coverage. Additionally, while third grade data indicated significant correlations between some measures of topic coverage and achievement, these relationships are not linked conceptually in terms of content. In other words, on some achievement subscales, coverage/duration of content other than that measured on the subscale was more strongly correlated with achievement than coverage/duration on the specific content measured by the subscale.

In contrast, fourth grade achievement correlated most highly and significantly with variables measuring instructional practices rather than topic coverage. The frequency with which fourth grade teachers report that their students practice computation in class correlates negatively with achievement across subscales (and significantly so for two subscales). On the other hand, the frequency with which fourth grade students are required to address problems without obvious methods of solution correlates positively and significantly with student achievement.

Finally, while a variety of significant correlations were found between different OTL measures and class achievement on subscale scores, the size of correlations were only small to moderate, with the strongest relationship between any variables at .459, between teacher reports about mathematics curricular focus and third grade student performance on the measurement subscale. In general, these data indicate a modest relationship between OTL and achievement as operationalized in the study.

Discussion

These findings have several implications for considering the promise of large-scale classroom practice measures for providing reliable information about the equity of student opportunities to learn. First of all, the data patterns described here indicate that a survey measure of OTL can provide some reasonable information about the nature and distribution of instructional practices in different classrooms. In terms of the first purpose described in the literature about OTL indicators–to monitor equity processes--this type of measure can provide useful information. It must be remembered, however, that this information is at a relatively "broad-brush" level of specificity. It primarily shows that different classrooms are characterized by different levels of exposure to content and instructional practices.

Perhaps the strongest example of variation in content exposure–and consequent questions about the equity of student experiences in the classroom–may be found in teacher reports about the fractions-related content that they cover. Table 4 provides a description of the proportions of teachers in the sample who report coverage of different specific fractions-related topics and significant differences in coverage by grade, indicating considerable articulation of content by grade level.

Examining the frequencies in Table 4 helps to clarify the distribution of content coverage across sampled classrooms–and it varies considerably across classrooms. For example, we can see that the most commonly covered fractions topic is also the most general–the topic addressing the meaning, representation and uses of common fractions. More than 60% of third grade classes and almost 80% of fourth grade classes had access to this content. However, focusing on fourth grade classes, more than one-fifth of the fourth grade students who took the TIMSS measure had no exposure to this material. Forty percent had no instruction about the operations of common fractions prior to the test, and almost 60% lacked exposure to information about the operations of decimal fractions. But there are other classrooms in the sample where students did receive instruction on these topics–even some third-grade classrooms. These patterns indicate that equity of opportunity is indeed an issue for these students.

Table 4. Teacher Topic Coverage on Fractions and Grade Level Differences

Topic	*Grade 3*		*Grade 4*		*Pearson Chi-Square*		
	N	*% covering topic*	*N*	*% covering topic*	*Value*	*df*	*2-tailed Sig.*
Common and decimal fractions **	47	29.8%	55	61.8%	10.437	1	.001
Meanings, representation and uses of decimal fractions**	47	19.1%	55	41.8%	6.049	1	.014
Operations of decimal fractions**	47	17%	55	41.8%	7.366	1	.007
Properties of decimal fractions**	47	14.9%	55	38.2%	6.901	1	.009
Meaning, representation and uses of common fractions	47	61.7%	55	78.2%	3.315	1	.069
Operations of common fractions*	47	38.3%	55	60%	4.774	1	.029
Properties of common fractions*	47	38.3%	55	63.6%	6.519	1	.011
Relationships between common & decimal fractions**	47	8.5%	55	53.9%	15.701	1	.000
Finding equivalent fractions and forms**	47	34%	55	67.3%	11.211	1	.001
Ordering of fractions (common and decimals)**	47	17%	55	49.1%	11.564	1	.001

* p<.05
** p<.01

When considered in conjunction with the information about the fractions content covered in these classes, data patterns measured by the curricular focus variable further support the overall picture of some classes that get access to more content topics than others do–and this is the case across grade levels. As illustrated in Table 5, for some third-grade classes, the same number or more topics (which includes a variety of ambitious, fractions-related topics) are covered as in some fourth-grade classes. One plausible explanation for such between-class variation may be related to student ability grouping practices, although it is not possible to verify this type of hypothesis through the TIMSS data.

Table 5. Descriptive statistics of curricular focus by grade

Grade Level	*Mathematics Topics Reported (of 36 possible)*				
	N	*Min.*	*Max.*	*M*	*S*
3	38	8	29	16.39	5.11
4	51	9	35	22.02	6.08

However, these data trends provide only a very rough, general sense about the distribution of learning opportunities across student classrooms. Further, they raise serious technical questions about how large-scale measures might be used within an accountability policy context and in connection with achievement. When we consider the second and third purposes for OTL indicators described in the literature–to inform accountability policies and particularly to connect OTL processes with student achievement-- the data patterns described here illustrate a few technical problems with the use of large-scale measures to examine classroom processes. These problems relate to requirements for quality related to the potential stakes of data use, issues related to the relation between OTL and the achievement measure (as well as the achievement measure's curricular validity at the school level), and the general limitations of survey measures.

A serious issue in the use of large-scale OTL measures within a high-stakes accountability context is that such a context exerts certain requirements for a measure's technical quality. Since technical requirements for measures become more stringent depending on the stakes associated with testing results (e.g., AERA, 1999, p. 139), and OTL describes the conditions necessary for students to learn what is expected on tests, by extension, quality requirements for OTL measures shift depending on the policy context and projected use of data. Therefore, large-scale measures of achievement and classroom processes need to have strong evidence of technical qualities, particularly as they may need to be legally defensible. Achievement measures must meet extremely stringent criteria for reliability and validity. Measures of OTL should be linked to specific achievement measures that have demonstrated proof of their sensitivity to OTL across established scales. Clearly, the TIMSS measure, as used in this study, cannot meet such criteria, regardless of how rigorously it may have been designed for the purposes of cross-national comparison.

These technical issues arise whenever a measure (of any sort) is used for purposes other than that for which it was expressly designed. Because of the nature of the TIMSS data set and its measures, which were designed for a certain purpose (e.g., cross-national or cross-state comparisons over a broad domain of data), I made certain adjustments and decisions about data treatment in order to examine internal data patterns between classroom processes and achievement. These included decisions related to 1) issues of sampling and weighting, 2) operationalizing achievement and OTL, and 3) the types of analyses to use and how to link achievement and OTL data. These are decision points that might arise in any secondary analyses of data–particularly in examining achievement and OTL data. And each of these decisions may involve manipulation of the data in ways inconsistent with the study's original purpose–thus affecting the quality and legal defensibility of the findings. Interestingly, using large-scale measures for purposes other than those for which they are expressly designed is not an uncommon practice; policy makers tend to emphasize the desirability of multiple uses for measures, contrary to the recommendations of measurement experts (McDonnell, 1994), which are made for exactly the reasons raised here.

Another key aspect of OTL that has traditionally been ignored is that, within the research, it is measure-specific. It addresses student opportunities to learn the content on which they will be tested *using a specific measure*. However, in policy conversations about standards in the early 1990's, OTL was conceptualized much more generically as the opportunity to learn the content standards *writ large*. A question for further research is that of OTL's utility as a research concept, given this broader policy definition. What is the balance that we, as researchers, can strike? If OTL is bound to specific achievement measures, can rigorous research that examines the OTL/achievement relationship provide useful information at a general enough level so that we can still build useful understandings about the context of successful classroom supports for learning independent of a particular achievement measure or a particular, highly specified definition of OTL? On the other hand, if we consider OTL to be assessment-free, how can we reconcile the relatively weaker OTL/achievement linkages, parse out the various sources of error, and still glean useful insights about how the nature and distribution of classroom processes relate to student learning?

A vital issue, particularly in connection to the idea of OTL as measure-bound, is that large-scale achievement measures may not be aligned with local achievement goals–that is, they may lack curricular

validity. The TIMSS items, for instance, do not represent a universal agreement about what should be taught in mathematics to 9-year old students, despite the fact that they have been rumored to be driving the curriculum in several African countries. Achievement items were selected from a cross-country pool, and were designed to be somewhat representative of countries' curricula at the national level, and to be "equally unfair" to all countries. (Wolf, 1998, p. 494). It is important to assess the content validity of any large-scale measure relative to local priorities and goals, rather than to accept large-scale ratings unquestioningly.

The key issue for local policy makers in selecting an achievement measure is that of alignment. To what extent does the content represented in *any* achievement measure correspond to the curriculum that is expected to be taught in local classrooms? While it may sound terribly obvious, it is important to say--*What the achievement test is and how well it is (or should be) aligned with local curriculum is important*. Psychometrically, there is evidence that the choice of achievement indicator matters considerably in deciding progress (Linn, 2000). What the measure of achievement actually *is* has considerably influence on whether we decide if progress is being made, and different measures may tell us quite different things about that progress, simply because it is being defined differentially. This is important for local policy makers to understand and articulate, particularly if external measures are being used to measure local progress and serious consequences are attached to results.

The final technical issue raised through this study involves the limitations of surveys and other large-scale measures to adequately capture the complexity of good classroom instruction–particularly ambitious classroom instruction. While most large-scale collection of OTL data has been through surveys because of the costs associated with classroom observations, the lack of standardized observation protocols, and the minimal generalizeability associated with such measures (Brewer & Stacz, 1996; Kennedy, 1999; McDonnell, 1995), surveys have limitations in what they can measure. On the one hand, we have some good information about what surveys of OTL can do; validity studies have indicated that teacher reports about current topic coverage are more reliable than reports about topics taught over the course of the year, that questions about instructional tools are less reliable than questions about content coverage, and that specific content questions are more reliable than more general items (Burstein, McDonnell, Van Winkle, Ormseth, Mirocha, & Guiton,1995; Yoon, Burstein, & Gold, 1991). On the other hand, surveys tend to be poor measures for assessing teacher expectations for students, particularly in light of reform-oriented perspectives (e.g., student-centered activity and construction of knowledge) and information about instructional strategies (Burstein, et al., 1995; Wiley & Yoon, 1995). To maximize the usefulness and quality of the information gleaned from surveys, it is necessary to supplement them with other data sources, such as teacher logs, student assignments, observations, or other classroom artifacts.

To summarize, large-scale measures of OTL have promise for providing general information about students' experiences in the classroom–and therefore, illuminating some equity issues in terms of the broad distribution of those experienced. However, they are not technically sound enough to provide rigorous data in an era of high-stakes accountability. As noted by Mayer (1999) and Spillane and Zeuli (1999) in validation studies, surveys can make some crude distinctions between teachers' practices, but teachers who look very much alike on surveys also look very different when observed, in terms of the nature and quality of their classroom discourse and the environment for student learning. According to Mayer, this indicates that "more work needs to be done to devise surveys that better distinguish between teachers who perfunctorily use reform practices and those who use them effectively." (p. 42) While surveys can capture the frequency with which teachers do a certain thing, they cannot yet capture the complexity of the teaching and learning process or even approximate a measure of whether a given frequency of teacher activity is, indeed, an appropriate thing to do. This indicates that, given the current policy environment, we need to re-think our approaches to large-scale measurement of OTL and to work toward the systematic development of broader indicator systems that take into account a nested design addressing aspects of OTL at multiple levels of the system, while also providing for more complex explorations into classroom practice.

References

American Educational Research Association (AERA) (1999). Standards for educational and psychological testing. Washington, DC: Author.

Blank, R. L. (1993). Developing a system of education indicators: Selecting, implementing, and reporting indicators. Educational Evaluation and Policy Analysis, 15 (1), 65-80.

Brewer, D. J., & Stacz, C. (1996). Enhancing opportunity to learn measures in NCES data. Santa Monica, CA: RAND.

Burstein, L., McDonnell, L., Van Winkle, J., Ormseth, T., Mirocha, J., & Guiton, G. (1995). Validating national curriculum indicators. Santa Monica, CA: RAND.

Campbell, P. B., & Clewell, B.C. (1999). Science, math, and girls...still a long way to go. Education Week, XIX, No. 2.

Cohen, D. K. (1990). Revolution in one classroom. In S. H. Fuhrman & B. J. Malen (Eds.) The politics of curriculum and testing (pp. 103-123). New York, NY: Falmer Press.

Education Week (January 11, 2001). A better balance: Standards, tests, and the tools to succeed. Special edition: Quality counts 2001.

Gamoran, A. (1987). The stratification of high school learning opportunities. Sociology of Education, 60, 135-155.

Goals 2000: Educate America Act, 20 U.S.C. §5801-6084 (1994).

Guiton, G., & Burstein, L. (1993, April). Indicators of curriculum and instruction. Paper presented at the Mini-conference on Educational Indicators at the Annual Conference of the American Educational Research Association, Atlanta, GA.

Guiton, G., & Oakes, J. (1995). Opportunity to learn and conceptions of educational equality. Educational Evaluation and Policy Analysis, 17 (3), 323-336.

Hanushek, E.A. (1997). Outcomes, incentives and beliefs: Reflections on analysis of the economics of schools. Educational Evaluation and Policy Analysis, 19 (4), 301-308.

Herman, J. L., Klein, D. C. D., & Wakai, S. T. (1996). Assessing equity in alternative assessment: An illustration of opportunity-to-learn issues. [CSE Technical Report 440]. Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Herman, J. L., & Klein, D. C. D. (1997). Assessing opportunity-to-learn: A California example. [CSE Technical Report 453]. Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Howe, K. R. (1989). In defense of outcomes-based conceptions of equal educational opportunity. Educational Theory, 39 (4), 317-336.

Kennedy, M. M. (1999). Approximations to indicators of student outcomes. Educational Evaluation and Policy Analysis, 21, (4), 345-364.

Kozol, J. (1991). Savage inequalities. New York: Crown Publishers.

Lawrenz, F. & Huffman, D. (1998a). Instructional relationships within the Third International Mathematics and Science Study (TIMSS) data for Minnesota third and fourth graders. University of Minnesota. Available: http://carei.coled.umn.edu.

Linn, R. L. (2000). Assessments and accountability. Educational Researcher, 29 (2), 4-15.

Martin, M. O. (1996). Third International Mathematics and Science Study: An overview, in M. O. Martin & D. L. Kelly (Eds.), Third International Mathematics and Science Study (TIMSS) technical report, Volume 1: Design and development. Chestnut Hill, MA: Boston College.

Mayer, D. P. (1999). Measuring instructional practice: Can policymakers trust survey data? Educational Evaluation and Policy Analysis, 21 (1), 29-45.

McDonnell, L. M. (1994). Policymakers' views of student assessment. Santa Monica, CA: RAND.

McDonnell, L. M. (1995). OTL as research concept. Educational Evaluation and Policy Analysis, 17, (3), 305-322.

McLaughlin, M. W., & Shepard, L. A. (1995). Improving education through standards-based reform. Stanford, CA: The National Academy of Education.

Muthen, B., Huang, L., Jo, B., Knoo, S., Goff, G.N., Novak, J.R., & Shih, J.C. (1995). Opportunity-to-learn effects on achievement: Analytical aspects. Educational Evaluation and Policy Analysis, 17, (3), 371-403.

National Council on Education Standards and Testing. (1992). Raising standards for American education. Washington, D.C.: U.S. Government Printing Office.

National Council of Teachers of Mathematics (NCTM). (1989). Curriculum and Evaluation Standards for School Mathematics. Reston, VA: Author.

Oakes, J. (1985). Keeping track: How schools structure inequality. New haven, CO: Yale University Press.

O'Day, J. A., & Smith, M.S. (1992). Systemic reform and educational opportunity. In S.H. Fuhrman (Ed.), Designing coherent education policy (pp. 250-312). San Francisco, CA: Jossey-Bass.

Porter, A. C. (1995). The uses and misuses of opportunity-to-learn standards. In D. Ravitch (Ed.) Debating the future of American education: Do we need national standards and assessments? (pp. 40-65). Washington, D.C.: The Brookings Institute.

Raizen, S. A., & Jones, L. J. (Eds.). (1985). Indicators of pre-college education in science and mathematics: A preliminary review. Washington, DC: National Academy Press.

Schmidt, W. H., & McKnight, C. C. (1995). Surveying educational opportunity in mathematics and science: An international perspective. Educational Evaluation and Policy Analysis, 17, (3), 337-354.

Schmidt, W. H., McKnight, C. C., & Raizen, S. A. (1997). A splintered vision: An investigation of U.S. science and mathematics education. Dordrecht; Boston: Kluwer Academic Publishers.

Schmidt, W. H., McKnight, C. C., Cogan, L.S., Jakwerth, P. M., & Houang, R. T. (1999). Facing the consequences: Using TIMSS for a closer look at U.S. mathematics and science education. Dordrecht; Boston: Kluwer Academic Publishers.

Spillane, J. P., & Zeuli, J. S. (1999). Reform and teaching: Exploring patterns of practice in the context of state and national reforms. Educational Evaluation and Policy Analysis, 21 (1), 1-27.

Stecher, B. (1992). Describing secondary curriculum in mathematics and science: Current status and future indicators (N03406-NSF) Santa Monica, CA: RAND.

Traiman, S. L. (1993). The debate on opportunity-to-learn standards. Washington, DC: National Governors' Association.

U.S. Department of Education, National Center for Education Statistics (1996). Pursuing excellence: A study of U. S. eighth-grade mathematics and science achievement in international context. (NCES 96-255). Washington, DC: U. S. Government Printing Office.

U.S. Department of Education, National Center for Education Statistics (1997). Pursuing excellence: A study of U. S. fourth-grade mathematics and science achievement in international context. (NCES 97-255). Washington, DC: U. S. Government Printing Office.

U.S. Department of Education, National Center for Education Statistics (1998). Pursuing excellence: A study of U. S. twelfth-grade mathematics and science achievement in international context. (NCES 98-049). Washington, DC: U. S. Government Printing Office.

Voelkl, K.E. (1998). Mathematics achievement in Colorado in the primary school years: Results from IEA's Third International Mathematics and Science Study. Princeton, NJ: Educational Testing Services.

Welner, K. G., & Oakes, J. (1996). (Li)Ability grouping: the new susceptibility of school tracking systems to legal challenges. Harvard Educational Review, 66 (3), 451-470.

WESTAT, Inc. (1997). User's guide for the Third International Mathematics and Science Study (TIMSS) and U.S. augmented data files--Draft, November 1997. Rockville, MD: Author.

Wiley, D. E., & Yoon, B. (1995). Teacher reports on opportunity to learn: Analyses of the 1993 California Learning Assessment System (CLAS). Educational Evaluation and Policy Analysis, 17 (3), 355-370.

Wolf, R. M. (1998). Validity issues in international assessments. International Journal of Educational Research, 29, 491-501.

Yoon, B., Burstein, L., & Gold, K. (1991). Assessing the content validity of teachers' reports of content coverage and its relationship to student achievement. In L. Burstein (Ed.) Models for validating content coverage. Instructional assessment project. Available ERIC ED 335367. Los Angeles, CA: National Center for Research on Standards, Evaluation, and Student Testing (CRESST).